



Performance Comparison

CLS Token:

- Learnable aggregation
- Task-specific
- BERT's approach

- ✓ Pre-trained
- ✓ Effective
- ✗ Single vector

Mean Pooling:

- Simple averaging
- No learning
- Dilutes information

- ✓ Simple
- ✓ No params
- ✗ Poor performance

Max Pooling:

- Sparse selection
- Loses context
- Task-dependent

- ✓ Sparse
- ✓ No params
- ✗ Loses nuance

Attention Pool:

- Flexible weighting
- More parameters
- Often competitive

- ✓ Adaptive
- ✓ Flexible
- ✗ More complex